D986–D990 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 16 October 2007*
doi:10.1093/nar/gkm757

CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform

Séverine Gagnot[1], Jean-Philippe Tamby[2], Marie-Laure Martin-Magniette[1,3], Frédérique Bitton[1], Ludivine Taconnat[1], Sandrine Balzergue[1], Sébastien Aubourg[1], Jean-Pierre Renou[1], Alain Lecharny[1,4] and Véronique Brunaud[1,*]

[1]Unité de Recherche en Génomique Végétale (URGV) - UMR INRA 1165-CNRS 8114-UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, [2]Laboratoire de Biologie Cellulaire - Institut J.P. Bourgin - INRA Centre de Versailles-Grignon, Route de Saint Cyr (RD 10), 78026 Versailles Cedex, France, [3]Unité de Mathématiques et Informatique Appliquées (MIA) - UMR 518 AgroParisTech-INRA, 16 Rue Claude Bernard, 75231 Paris Cedex and [4]Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP) - UMR CNRS UPS Bâtiment 630, 91405 Orsay Cedex, France

Received July 12, 2007; Revised September 7, 2007; Accepted September 11, 2007

ABSTRACT

CATdb is a free resource available at http://urgv.evry.inra.fr/CATdb that provides public access to a large collection of transcriptome data for *Arabidopsis thaliana* produced by a single Complete Arabidopsis Transcriptome Micro Array (CATMA) platform. CATMA probes consist of gene-specific sequence tags (GSTs) of 150–500 bp. The v2 version of CATMA contains 24 576 GST probes representing most of the predicted *A. thaliana* genes, and 615 probes tiling the chloroplastic and mitochondrial genomes. Data in CATdb are entirely processed with the same standardized protocol, from microarray printing to data analyses. CATdb contains the results of 53 projects including 1724 hybridized samples distributed between 13 different organs, 49 different developmental conditions, 45 mutants and 63 environmental conditions. All the data contained in CATdb can be downloaded from the web site and subsets of data can be sorted out and displayed either by keywords, by experiments, genes or lists of genes up to 100. CATdb gives an easy access to the complete description of experiments with a picture of the experiment design.

INTRODUCTION

Transcriptome characterization by microarray technologies is a powerful tool for functional analysis of genes. The primary purpose of most of the experiments was finding candidate genes for further experimental work. Nevertheless, with the accumulation of data, a complementary usage of the transcriptome resource is the integration of large sets of data to infer, for instance, gene regulatory networks. Several databases dedicated to microarray data exist and can be distributed in three general classes (i) public repositories including ArrayExpress, Gene Expression Omnibus (GEO) and The Center for Information Biology Gene Expression Database (CIBEX) (1–3); (ii) general databases oriented toward tools for the analyses and displaying of different types of arrays, like Genevestigator or the Stanford Microarray Database (SMD) (4,5) and (iii) specific databases dedicated to a species like The Arabidopsis Information Resource, or the expression browser (eFP) from the Bio-Array Resource for Arabidopsis Functional Genomics, or specific to a life kingdom like the Plant Expression Database (PLEXdb) (6–8). Despite considerable and valuable efforts done to define and apply the Minimal Information About Microarray Experiment (MIAME) (9) recommendations, a recent survey of the data in public repositories indicated that data submission and quality are troublesome for integrating current microarray data (10). The diversity of transcriptome data and methods to analyse them is one of the problems for the occasional users. We have developed CATdb to manage the microarray data resource generated by the URGV transcriptome platform (http://www.versailles.inra.fr/urgv) and allow an easy access to the data by the community of biologists. We took advantage of the unique origin of the URGV-CATMA data to concentrate our effort on the quality of the data and to systematically collect a global view of each project with the details of the experiment design. Thus, CATdb provides an easy

*To whom correspondence should be addressed. Tel: +33 1 60 87 45 14; Fax: +33 1 60 87 45 49; Email: brunaud@evry.inra.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

access to a large and growing set of microarrays named CATMA (Complete Arabidopsis Transcriptome Micro Array) (11). All the RNA samples are sent by collaborators at the URGV then checked for quality control, labelled and hybridized following normalized protocols. The scanning is performed with common settings and a unique normalization followed by a statistical analysis procedure is applied to each experiment as described subsequently.

CATMA MICROARRAYS

CATMA is a generic *Arabidopsis thaliana* microarray developed by a European consortium (12). The design of the probes for CATMA microarrays is different from the design of both the *A. thaliana* Agilent arrays (Palo Alto, CA, USA) and the ATH1 Affymetrix GeneChips (Santa Clara, CA, USA) (13) that use respectively oligo-nucleotide probes of 60 mers and sets of oligo-nucleotides of 25 mers. CATMA probes consist of gene-specific sequence tags (GSTs) of 150–500 bp that have been designed with SPADS (Specific Primers & Amplicons Design Software) (14). Tagged genes come from both the EuGene software prediction (15) and the annotation from The Institute for Genomic Research (TIGR).

The v2 version of CATMA contains 24 576 GST probes representing ~85% of the predicted genes, 615 probes tiling the chloroplastic and mitochondrial genomes (v2.1) and 44 probes of non-protein coding genes (v2.2). A thorough benchmark study established the CATMA array as a mature alternative to the Affymetrix and Agilent platforms (16). The CATMA GSTs are also the basic materials in the AGRIKOLA (Arabidopsis Genomic RNAi Knock-out Line Analysis) European project focusing on the large-scale systematic RNAi silencing of *Arabidopsis* genes (http://www.agrikola.org/).

DATABASE CHARACTERISTICS AND CONTENTS

Primarily, CATdb was based on the schema and objects used in the ArrayExpress database (17). Then, the ArrayExpress schema has been adapted to our platform to include some new features. The main differences are: (i) the systematic addition of a figure describing the design of an experiment in standardized format, (ii) the possibility to manage a supplementary step with the pooling of samples or extracts and (iii) the storage of the statistical analyses using technical replicates (see the Data Analysis section).

The complete description of the experiments is submitted via a private web interface that helps to respect the MIAME instructions. CATdb generates the SOFT (Simple Omnibus Format in Text) format developed by the GEO repository (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/). Data from the 1724 hybridizations in CATdb are also available either at GEO or at ArrayExpress. In the description of each project or experiment, there is the corresponding access number in GEO or ArrayExpress with a link to their respective web pages.

All the data submissions and analyses are performed in our laboratory, so the development of CATdb has been oriented by the visual approaches used by biologists at URGV, like using colours to encode the data values, to facilitate the analyses and comparisons of the data.

The increasing number of research projects involving CATMA microarrays shows that the CATMA arrays are an important tool for biologists. CATdb gives a public access to all the data produced by the URGV-CATMA platform even those that have not been published after one year. CATdb contains 46 projects with 1724 hybridizations corresponding to 627 different samples. The samples of these projects concern 13 types of organs: cells (73 samples), protoplasts (10), roots (116), hypocotyls (24), stems (18), leaves (129), flowers (36), pollen (2), siliques (4), seeds (17), whole aerial plants (43), plantlets (39) or whole plants (116). These samples are distributed between 49 different developmental conditions, 31 developmental stages, 45 mutants and 63 different abiotic/biotic stresses or treatments.

DATA ACCESS

CATdb is a free web resource available at the following address: http://urgv.evry.inra.fr/CATdb. There are four different possibilities to select a subset of data. First, a list of all the available projects is displayed by default. A limited list may be obtained by querying the database by keywords. These keywords are searched for in both the description of the projects, i.e. coordinator name, experiment type, environmental or treatment factor, mutant name and the description of the samples, i.e. plant species, organs, treatments and type of arrays. Second, an experiment name may be selected in the project table giving access to the entire description of the corresponding experiment including a picture of the experiment design (Figure 1A). The swap column gives access to all the results of hybridizations organized by dye-swap for the selected experiment. Normalized $\log_2$ intensities, $\log_2$-ratios and Bonferroni *P*-values are given for each probe (Figure 1B). As this table is rather large, only the probes with statistically significant differential expression for a dye-swap are displayed on the screen. Nevertheless, the complete table may be downloaded as a tabulated text file. Third, from either a gene or a probe accession, one may obtain signal intensities and Bonferroni *P*-value for all the dye-swaps processed in all the projects (Figure 2). Furthermore, data may be sorted by project, organ or any statistics. For each probe, the associated features, i.e. sequence, quality of PCR results and if applicable, the tagged gene with functional annotation, are given. Fourth, from a list of genes or probe accessions, one obtains a table containing, for each selected probe, the $\log_2$-ratios for all the projects (Figure 3). The coloured display of the differential expression allows the comparison of the data for a list of up to 100 genes.

All the public data contained in CATdb can be downloaded from an anonymous FTP (File Transfer Protocol) site (ftp://urgv.evry.inra.fr/CATdb). Users who

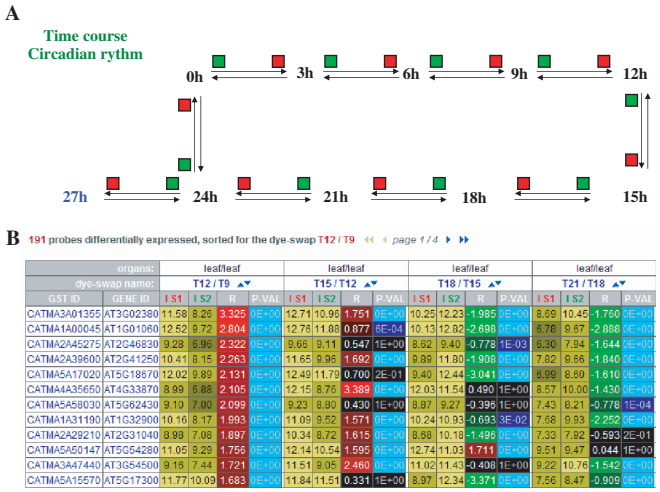

Figure 1. Results of a query of CATdb for an experiment called ''Circadian cycle' and belonging to the project named 'AF30_Starch_circadian_rythm'. The result includes (**A**) the experimental design that describes all the hybridizations with two colours, red and green, indicating the dye used for labelling each sample and (**B**) a table displaying for each dye-swap of the experiment and from left to right, the $\log_2$-intensities for samples 1 and 2, the $\log_2$-ratio and the Bonferroni *P*-value. In the example, results are sorted by the ratio values in the dye-swap between the leaf sample T12, extracted at 12 h after the start of the experiment, and the leaf sample T9, extracted at 9 h.

have subscribed to the CATdb e-mailing list receive news about updates and new tools.

DATA ANALYSIS

Statistic methods were developed under the software R (R Development Core Team, http://www.R-project.org) in collaboration with the group 'Statistics and Genome' at UMR AgroParisTech/INRA MIA 518 and are available in the R package 'Anapuce' on their web site (http://www.inapg.fr/ens_rech/maths/outil_A.html). For each CATMA array, the raw data include the logarithm of median feature pixel intensity at wavelengths 635 nm (red) and 532 nm (green), no background is subtracted. A normalization per array is performed to remove systematic biases. First, spots that are considered badly formed features are excluded. Then, a global intensity-dependent normalization is performed using the lowess procedures (18) to correct the dye bias. Finally, for each block, the log-ratio median calculated over the values for the entire block is subtracted from each individual log-ratio value to correct effects on each block, as well as print-tip, washing and/or drying effects. At the end of the normalization step, a normalized log-ratio, which is equivalent to an expression difference (in log base 2) between the two samples co-hybridized on the same array,

PROJECT	EXPERIMENT	ARRAY TYPE	ORGANS	S1 / S2	I S1	I S2	R= IS1-IS2	P-VAL
AF24_CytP450_Cyp98A3	polyphenol	CATMA	whole plant	Tn4 / WS3	12.80	9.02	3.776	0E+00
			whole plant	Tn4 / WS3	11.35	8.13	3.220	0E+00
AF30_Starch_circadian_rythm	Circadian cycle	CATMA	leaf	T24 / T0	6.38	7.11	-0.732	1E-03
			leaf	T21 / T18	7.68	9.93	-2.252	0E+00
			leaf	T27 / T24	6.76	6.92	-0.164	1E+00
			leaf	T18 / T15	10.24	10.93	-0.693	3E-02
			leaf	T21 / T24	8.30	6.73	1.575	0E+00
			leaf	T12 / T9	10.16	8.17	1.993	0E+00
			leaf	T15 / T12	11.09	9.52	1.571	0E+00
			leaf	T6 / T3	6.64	6.45	0.185	1E+00
			leaf	T3 / T0	6.98	7.56	-0.587	1E+00
			leaf	T9 / T6	7.88	6.74	1.134	0E+00
AF30_Starch_circadian_rythm	Double mutant SBE	CATMA	leaf	sbe1_sbe2 nd 1 / WS nd 1	10.42	12.34	-1.923	0E+00
			leaf	sbe1_sbe2 nd 3 / WS nd 3	10.49	12.80	-2.314	0E+00
			leaf	sbe1_sbe2 dn 1 / WS dn 1	8.36	8.69	-0.329	1E+00
			leaf	sbe1_sbe2 dn 2 / WS dn 2	8.33	10.96	-2.632	0E+00
			leaf	sbe1_sbe2 dn 3 / WS dn 3	8.62	11.74	-3.128	0E+00
			leaf	sbe1_sbe2 nd 2 / WS nd 2	10.21	12.10	-1.889	0E+00
AF31_HAF2_GCN5	Comparison WS vs HAF2_GCN5	CATMA	plantlet	GCN5_repeat1 / WS_repeat1	10.14	11.76	-1.617	0E+00
			plantlet	GCN5_Bio / WS_Bio	8.89	9.84	-0.946	9E-03
			plantlet	GCN5_repeat2 / WS_repeat2	10.63	12.13	-1.507	0E+00
			plantlet	HAF2_Bio / WS_Bio	9.26	9.61	-0.348	1E+00
			plantlet	HAF2_repeat1 / WS_repeat1	10.19	11.77	-1.581	0E+00
			plantlet	HAF2_repeat2 / WS_repeat2	10.05	11.61	-1.559	0E+00

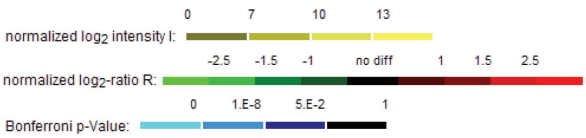

Figure 2. Results of a query of CATdb with the name of a gene. The query was the gene AT1G32900. Data are displayed for all the projects in CATdb, but only the first four projects sorted by the Bonferroni *P*-values are shown here. From the left to the right, columns contain the project and the experiment names, the array type, the organ used, the sample name, the $\log_2$-intensities for both samples, the $\log_2$-ratio and the Bonferroni *P*-value.

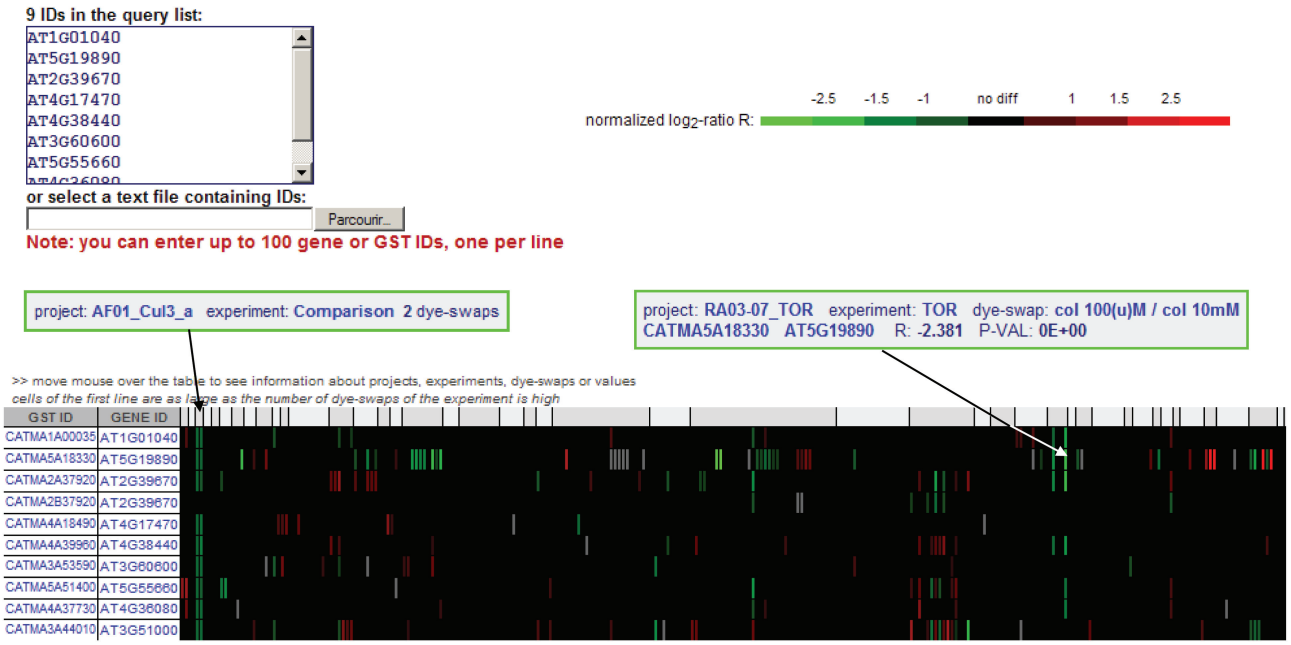

Figure 3. Results of a query of CATdb with a list of gene identifiers. From the left to the right, columns are the probe name, the corresponding gene name, the differential expression $\log_2$-ratio in all the different projects contained in CATdb. The differential expression is colour coded in green, red or black corresponding to the levels of $\log_2$-ratio in each swap analysed, and in grey for missing values. The correspondence between the colours and the $\log_2$-ratio values is given in a bar above the table. By rolling over a cell of the table, as shown by the arrows, one may display more information about either the projects, dye-swaps or expression values, depending on the selected cell.

is given for each spot. It is equal to the raw log-ratio minus the lowess correction minus the block correction. A normalized logarithm intensity for each sample is also calculated. It is done according to the within-array correction proposed by Yang and Thorne (19), which is a redistribution of the correction calculated for the $\log_2$-ratio normalization on each channel.

To determine differentially expressed genes from a dye-swap, a paired *t*-test is performed on the $\log_2$-ratios. Since the number of observations per spot equals two, it is inadequate for calculating a specific variance. For this reason, it is assumed that the variance of the $\log_2$-ratios is the same for all spots. This solution has the main advantage to calculate an estimator over a large number of data, leading to a robust estimation of the variance and to a gain in the power of the test. Nevertheless, this solution should be applied with some precautions since some spots display an extreme specific variance (too small or too large) and prevent that the assumption of common variance is verified. Indeed spots with a too small specific variance decrease wrongly the estimate of the common variance and hence it could lead to increase the number of false positives, and spot with a too large variance increase wrongly the estimate of the common variance and hence it could lead to decrease the test power. For the above reasons, spots with extreme specific variance are excluded from the statistical analysis. The spots that are excluded are those with a 'specific variance/common variance' ratio smaller than the 'alpha-quantile of a chi-squared distribution of one degree of liberty' or greater than the '1-alpha-quantile of a chi-squared distribution of one degree of liberty' with alpha equal to 0.0001. This rule stems from a direct application of Cochran's theorem. The raw *P*-values are adjusted by the Bonferroni method, which controls the Family Wise Error Rate (FWER) (20). When the Bonferroni *P*-value is lower than 0.05, the spot is declared differentially expressed. Spots with a missing *P*-value are spots with an extreme variance or genes for which one observation only is available. That is, when for one of the two arrays, the spot corresponding to the gene was a badly formed feature.

DATA QUALITY

Information on the CATMA probes and the corresponding genes is available in CATdb. This includes probe sequences and their estimated specificity, amplification efficiency, localization within genes (intron, exon) or between genes. All these annotations are graphically displayed in the genome database FLAGdb^{++} (21) and there are direct links from probes and genes in CATdb toward the probe loci in FLAGdb^{++}. To validate transcriptome data, the biologist relies on quantitative RT-PCR applied to a set of genes exhibiting differential expression between two experimental situations. On the CATMA resource, quantitative RT-PCRs were done on more than 200 genes and CATMA results have been confirmed in more than 90% of the validations. The details of RT-PCR from tested genes are described in the publications associated to the different research projects using CATMA arrays. A list of these publications is available on the CATdb web site.

FUTURE PLANS

Based on the number of not yet public data, 4336 hybridized samples, stored in CATdb, the number of public projects is expected to double in the coming year. Updating data depends on the submission date of a project. As in most public repositories, data cannot be maintained under the private status more than one year and any data are publicly released after this period of time or before on the authors' request.

CATMA is an ongoing project and new array designs will be released soon including 7189 new GSTs (collaboration with CATMA members) tagging the remaining annotated genes and different paralogues belonging to a gene family. Furthermore, probes for small RNA genes were designed by URGV in collaboration with O. Voinnet and L. Navarro (IBMP Strasbourg) and will be included in a future version. CATdb developments needed by the new designs are done in parallel.

ACKNOWLEDGEMENTS

The authors would like to thank Franck Samson for his help during the optimization of the code and Philippe Grevet for assistance and management of the ftp and web servers. Funding to pay the Open Access publication charges for this article was provided by INRA. The URGV CATMA resource has been funded by Génoplante.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P. *et al.* (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
2. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. *et al.* (2006) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
3. Ikeo,K., Ishi-I,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol.*, **326**, 1079–1082.
4. Zimmermann,P., Hennig,L. and Gruissem,W. (2005) Gene expression analysis and network discovery using Genevestigator. *Trends Plant Sci.*, **10**, 407–409.
5. Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F. *et al.* (2007) The Stanford MicroArray Datbase: implementation of new analysis and open source release of software. *Nucleic Acids Res.*, **35**, D766–D770.
6. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
7. Schmid,M., Davison,T.S., Henz,S.R., Pape,U.J., Demar,M., Vingron,M., Schölkopf,B., Weigel,D. and Lohmann,J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.
8. Shen,L., Gong,J., Caldo,R.A., Nettleton,D., Cook,D., Wise,R.P. and Dickerson,J.A. (2005) BarleyBase – An expression profiling database for plant genomics. *Nucleic Acids Res.*, **33**, D614–D618.
9. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
10. Larsson,O. and Sandberg,R. (2006) Lack of correct data format and comparability limits future integrative microarray research. *Nat. Biotechnol.*, **24**, 1322–1323.
11. Crowe,M.L., Serizet,C., Thareau,V., Aubourg,S., Rouzé,P., Beynon,J.L., Hilson,P., Weisbeek,P., Van Hummelen,P. *et al.* (2003) CATMA – A complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.
12. Hilson,P., Allemeersch,J., Altmann,T., Aubourg,S., Avon,A., Beynon,J., Bhalerao,R., Bitton,F., Caboche,M. *et al.* (2004) Versatile gene-specific sequence tags for arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res.*, **14**, 2176–2189.
13. Redman,J.C., Haas,B.J., Tanimoto,G. and Town,C.D. (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J.*, **38**, 545–561.
14. Thareau,V., Déhais,P., Serizet,C., Hilson,P., Rouzé,P. and Aubourg,S. (2003) Automatic design of gene-specific sequence tags for genome-wide functional studies. *Bioinformatics*, **19**, 2191–2198.
15. Schiex,T., Moisan,A. and Rouzé,P. (2001) EuGene: an eukaryotic gene finder that combines several sources of evidence. *Lect. Notes Comput. Sci.*, **2066**, 111–125.
16. Allemeersch,J., Durinck,S., Vanderhaeghen,R., Alard,P., Maes,R., Seeuws,K., Bogaert,T., Coddens,K., Deschouwer,K. *et al.* (2005) Benchmarking the CATMA microarray: a novel tool for Arabidopsis transcriptome analysis. *Plant Physiol.*, **137**, 588–601.
17. Brazma,A., Sarkans,U., Robinson,A., Vilo,J., Vingron,M., Hoheisel,J. and Fellenberg,K. (2002) Microarray data representation, annotation and storage. *Adv. Biochem. Eng. Biotechnol.*, **77**, 113–139.
18. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
19. Yang,Y.H. and Thorne,N.P. (2003) Normalization for two-color cDNA microarray data. In Goldstein,D. R. (ed.), *Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes–Monograph Series*, Vol. 40, pp. 403–418.
20. Ge,Y., Dudoit,S. and Speed,T.P. (2003) Resampling-based multiple testing for microarray data analysis. *TEST*, **12**, 1–44.
21. Samson,F., Brunaud,V., Duchene,S., De Oliveira,Y., Caboche,M., Lecharny,A. and Aubourg,S. (2004) FLAGdb^{++}: a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Res.*, **32**, D347–D350.